AD-A065 195 BROWN UNIV PROVIDENCE R I DIV OF APPLIED MATHEMATICS F/G 12/1

RATES OF CONVERGENCE FOR STOCHASTIC APPROXIMATION TYPE ALGORITH--ETC(U)

OCT 78 H J KUSHNER, H HUANG AFOSR-76-3063

UNCLASSIFIED AFOSR-TR-79-0084 NL

1 OF 1

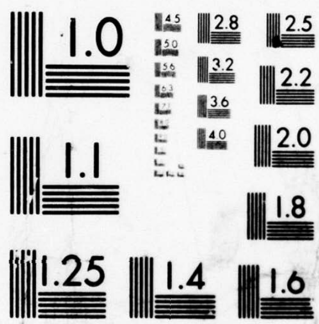
AD A065195

END

DATE FILMED

4--79

DDC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

18 19

LEVEL II

2

AD A065195

DDC FILE COPY

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR-TR- 79-0084

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)

6 RATES OF CONVERGENCE FOR STOCHASTIC APPROXIMATION TYPE ALGORITHMS.

5. TYPE OF REPORT & PERIOD COVERED

9 Interim rept.

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

10 Harold J. Kushner [redacted] Hai Huang

8. CONTRACT OR GRANT NUMBER(s)

15 AFOSR-76-3063

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Brown University
Division of Applied Mathematics
Providence, Rhode Island 02912

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

61102F 16 2304 A1 17 A1

11. CONTROLLING OFFICE NAME AND ADDRESS

Air Force Office of Scientific Research/NM
Bolling AFB, Washington, DC 20332

12. REPORT DATE

11 Oct[redacted]78

13. NUMBER OF PAGES

20 12 22 p.

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)

UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We consider the general form of the stochastic approximation algorithm $X_{n+1} = X_n + a_n h(X_n,\xi_n)$, where h is not necessarily adaptive in ξ_n. Such algorithms occur frequently in applications to adaptive control and identification problems, where $\{\xi_n\}$ is usually obtained from measurements of the input and output, and is

DD FORM 1 JAN 73 1473

065 300

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract continued

almost always complicated enough that the more classical assumptions on the noise fail to hold. Let $a_n = A/(n+1)$, $0 < \alpha < 1$, and let $X_n \to \theta$ w.p.1. Define $U_n = (n+1)^{\alpha/2}(X_n - \alpha)$. Then, loosely speaking, it is shown that the sequence of suitable continuous parameter interpolations of the sequence of "tails" of $\{U_n\}$ converges weakly to a Gaussian diffusion. From this we can get the asymptotic variance of U_n as well as other information. The assumptions on $\{\xi_n\}$ and $h(\cdot,\cdot)$ are quite reasonable from the point of view of applications.

(2)

AFOSR-TR- 79-0084

RATES OF CONVERGENCE FOR STOCHASTIC APPROXIMATION TYPE ALGORITHMS

Harold J. Kushner* and Hai Huang**

October, 1978

Abstract

We consider the general form of the stochastic approximation algorithm $X_{n+1} = X_n + a_n h(X_n, \xi_n)$, where h is not necessarily additive in ξ_n. Such algorithms occur frequently in applications to adaptive control and identification problems, where $\{\xi_n\}$ is usually obtained from measurements of the input and output, and is almost always complicated enough that the more classical assumptions on the noise fail to hold. Let $a_n = A/(n+1)^\alpha$, $0<\alpha<1$, and let $X_n \to \theta$ w.p. 1. Define $U_n = (n+1)^{\alpha/2}(X_n-\theta)$. Then, loosely speaking, it is shown that the sequence of suitable continuous parameter interpolations of the sequence of "tails" of $\{U_n\}$ converges weakly to a Gaussian diffusion. From this we can get the asymptotic variance of U_n as well as other information. The assumptions on $\{\xi_n\}$ and $h(\cdot,\cdot)$ are quite reasonable from the point of view of applications.

*Brown University, Divisions of Engineering and Applied Mathematics (Center for Dynamical Systems).

Research supported in part by the Air Force Office of Scientific Research under AF-AFOSR 77-3063, in part by the National Science Foundation under NSF-Eng-77-12946 and in part by the Office of Naval Research under N0014-76-C-0279-P0002.

**Brown University, Division of Applied Mathematics (Center for Dynamical Systems).

Research supported by the National Science Foundation under NSF-Eng-77-12946.

Introduction

Rates of convergence for stochastic approximation problems were given in [1, 2, 3, 4], the latter two references getting better results via weak convergence methods, for both constrained and unconstrained systems.

A form of stochastic approximation algorithm which is of increasing importance is the following. Let $\{a_n\}$ denote a sequence of positive real numbers with $\sum_n a_n = \infty$, h a suitable function and $\{\xi_n\}$ a sequence of random variables. Define the sequence of R^r-valued random variables $\{X_n\}$ by

$$(1.1) \qquad X_{n+1} = X_n + a_n h(X_n, \xi_n).$$

In [1] - [4], the function h was essentially additive in ξ_n, as is usually the case in classical Kiefer-Wolfowitz and Robbins-Munro type stochastic approximation algorithms. Of course, if $\{\xi_n\}$ is a sequence of independent random variables, then $h(X_n, \xi_n)$ can be written in the form $E[h(X_n, \xi_n)|X_n] + \psi_n$, where $\psi_n = h(X_n, \xi_n) - E[h(X_n, \xi_n)|X_n]$ is a member of an orthogonal sequence, and we are back to the classical case. In the applications that we have in mind the $\{\xi_n\}$ can be rather general processes.

The more general form (1.1) arises in applications to problems in the recursive identification of the parameters of linear systems, or in the so-called self-tuning regulators or in other applications of adaptive systems [5, 6]. Such applications are the motivation for this work. Often X_n is an estimate of the vector system parameter and ξ_n is a random vector which is related to the measured inputs and outputs of the system. The rate of convergence problem for such situations has not been dealt with, and somewhat different methods are required.

In this paper we develop rate of convergence results for (1.1) under quite reasonable conditions. Owing to the way in which (1.1) arises in applications, the $\{\xi_n\}$ is rarely a sequence of independent random variables, and $E(h(X_n, \xi_n)|\xi_0, \ldots, \xi_{n-1})$ is rarely a function only of X_{n-1}. Thus classical rate of convergence methods (as in [1], [2]) cannot be used directly. We use some of the ideas in [3], [4], but adapted to our case, and under weaker conditions on the noise sequences.

The problem is formulated and some assumptions given in Section 2. Weak convergence of a sequence of normalized $\{X_n\}$ is given in Section 3, and the general rate result appears in Section 4.

2. Terminology and Problem Formulation

For $\alpha \in (0,1]$ and A a matrix, set $a_n = A/(n+1)^{\alpha}$. Since we are concerned with rates of convergence, we assume convergence (see [4] for a detailed discussion of the convergence both w.p. 1 and weakly). In particular, we suppose that there is a $\theta \in R^r$ such that $X_n \to \theta$ w.p. 1. Set $U_n = (n+1)^{\alpha/2}(X_n-\theta)$, $\Delta t_n = (n+1)^{-\alpha}$, $h_n = h(\theta,\xi_n)$ and $\bar{h}_n = (n+2/n+1)^{\alpha/2}h_n$. Let $h(\cdot,\xi)$ be continuously differentiable for each ξ, with the gradient $h_x(\cdot,\cdot)$ being Borel-measurable.

There is a function $O(\cdot)$ such that with H_n defined by (2.1), (2.2) holds. (See [3], eqn. (5.2) for a related calculation for the case where h is additive in ξ.)

$$(2.1)\qquad H_n = Ah_x(\theta,\xi_n) + \frac{\alpha}{2(n+1)^{1-\alpha}} I + O(\frac{1}{n+1})I$$
$$+ A(\frac{n+2}{n+1})^{\alpha/2} \int_0^1 [h_x(\theta+t(X_n-\theta),\xi_n)-h_x(\theta,\xi_n)]dt$$
$$+ A[(\frac{n+2}{n+1})^{\alpha/2} - 1]\, h_x(\theta,\xi_n)$$

(2.2)* $$U_{n+1} = (I + \Delta t_n H_n)U_n + A\sqrt{\Delta t_n}\,\bar{h}_n$$

For future use define $\delta W_n = \sqrt{\Delta t_n}\, h_n$, $\delta\bar{W}_n = \sqrt{\Delta t_n}\,\bar{h}_n$.

Lemmas 1 and 2 contain some preparatory results concerning the iteration (2.2), and tightness of $\{U_n\}$ (i.e., $\sup_n P(|U_n| \geq N) \to 0$ as $N \to \infty$) is proved in Theorem 1.

Next, following the general approach of [3], a sequence of processes $\{U^N(\cdot)\}$ is defined as follows. Let $t_n = \sum_{i=0}^{n-1} \Delta t_i$, $t_0 = 0$ and define $m(t) = \max\{k: t_k \leq t\}$. Set $U^N(0) = U_N$ and $U^N(t) = U_{N+n}$ in $[t_{N+n}, t_{N+n+1})$. Thus $U^N(\cdot)$ is a process whose paths are piecewise constant and in $D^r[0,\infty)$, the space of R^r-valued functions which are right continuous on $[0,\infty)$ and have left-hand limits on $(0,\infty)$. Since it will be important for us to go back and forth between the $\{U_n\}$ and $\{U^N(\cdot)\}$ sequences, the functions $m(\cdot)$ and t_n will be used quite frequently, occasionally (and regrettably) causing some complicated notation.

Owing to the scale factor $a_n = A\Delta t_n$, the interpolation $U^N(\cdot)$ is quite natural for this problem. In Theorem 2 it will be shown that $\{U^N(\cdot)\}$ is tight in $D^r[0,\infty)$ and converges weakly to the stationary linear Gaussian diffusion (4.1). As is common in applications of weak convergence theory, if a sequence of measures $\{u_n\}$ is tight and converges weakly to u (all on R^r or $D^r[0,\infty)$), and u_n and u are induced by processes $X^n(\cdot)$ and $X(\cdot)$, resp. (with paths in R^r or $D^r[0,\infty)$), then we abuse terminology and say that $\{X^n\}$ is tight and converges weakly to X. This weak convergence gives us the basic rate of convergence result. Some advantages of our approach are discussed in [3]. It yields the convergence in distribution (to a normally distributed random variable, the stationary distribution of (4.1)) of

---

*From (2.1) we can guess that if $\alpha = 1$ (resp. $\alpha < 1$) the "effective" component of H_n is $(Ah_x(\theta,\xi_n) + I/2)$ $(Ah_x(\theta,\xi_n)$, resp.).

$\{U_n\}$, but also more, since it gives information on the correlation structure of the process $\{U_{N+n}, n \geq 0\}$ for large N.

Remark on weak convergence. Billingsley [7] is the most comprehensive reference. The space D[0,T] is discussed in [7], Sections 14 and 15. A brief summary of relevant facts is given in [4], Chapter 2. $D^r[0,\infty)$ is endowed with the usual ([7], Section 14) Skorokhod topology, with which it is a complete separable metric space. Convergence in $D^r[0,\infty)$ occurs if, for some sequence $T \to \infty$, it occurs (for the truncated functions) in each $D^r[0,T]$.

Assumptions. (A1) - (A5) will be used throughout the paper.

(A1) $X_n \to \theta$ w.p. 1

(A2) $h(\cdot,\cdot)$ is a Borel function, continuously differentiable in its first argument for each value of the second, and the gradient $h_x(\cdot,\cdot)$ is Borel. Also $Eh(\theta,\xi_n) \equiv 0$ and

$$\int_0^1 [h_x(\theta+t(X_n-\theta),\xi_n)-h_x(\theta,\xi_n)]dt \to 0 \quad \text{w.p.1}$$

as $n \to \infty$. (Certainly true if the ξ_n are bounded and $h_x(\cdot,\cdot)$ is continuous.)

(A3a) There is a matrix H such that for some (hence each) $T > 0$ and each $\varepsilon > 0$

$$\lim_{n\to\infty} P\{\sup_{j\geq n} \max_{0\leq t\leq T} \left| \sum_{i=m(jT)}^{m(jT+t)-1} \Delta t_i(h_x(\theta,\xi_j)-H)\right| \geq \varepsilon\} = 0.$$

(A3b) There is a constant τ such that for each $\varepsilon > 0$ and $T > 0$,

$$\lim_{n\to\infty} P\{\sup_{j\geq n} \max_{0\leq t\leq T} \left| \sum_{i=m(jT)}^{m(jT+t)-1} \Delta t_i(|h_x(\theta,\xi_i)| - \tau)\right| \geq \varepsilon\} = 0,$$

where $|x| = (x'x)^{1/2}$ and $|M| = \sup_{|x|=1}|Mx|$ if M is a matrix.

Remark on (A3a and b). Conditions of type (A3a, A3b) were used extensively in the monograph [4], and as shown in that reference are rather weak and quite natural for the problem. See, for example, the several cases discussed in [ ], Chapter 2.2. The conditions are commonly satisfied by the noise processes which appear in the usual applications to the identification problem. We mention only the following three cases for (A3a): (a) $\sum a_n^2 < \infty$ and $\{h_x(\theta,\xi_n)-Eh_x(\theta,\xi_n)\}$ orthogonal; (b) $h_x(\theta,\xi_n)-Eh_x(\theta,\xi_n) = \sum_{j=0}^{\infty} b_j\psi_{n-j}$, for a broad class of $\{b_j\}$, $\{\psi_j\}$ where $\{\psi_j\}$ are independent and identically distributed; (c) $\{\xi_n\}$ stationary, (A5) holds for h_x replacing h and $\sum a_i^2(\log_2 i)^2 < \infty$ holds.

In order to illustrate our terminology and get some additional insight into (A3), let us define a process $\eta(t)$ as follows: $\eta(0) = 0$, and $\eta(t) = \sum_{i=0}^{n-1} \Delta t_i(h_x(\theta,\xi_i)-H)$ on $[t_n,t_{n+1})$. Then

$$\eta(t) = \sum_{i=0}^{m(t)-1} \Delta t_i(h_x(\theta,\xi_i)-H).$$

Condition (A3a) implies that the variation of the "increasing compressed interpolation" $\eta(t)$ over an arbitrary interval $(\alpha,\alpha+T)$ goes to zero w.p. 1 as $\alpha \to \infty$.

(A4) If $\alpha = 1$, set $\bar{H} = AH + I/2$, and if $\alpha < 1$, set $\bar{H} = AH$. The eigenvalues of $\bar{H}$ have negative real parts.

(A5) Define R_{mk} by $R_{mk} = Eh'(\theta,\xi_m)h(\theta,\xi_k)$. Then $\sup_m \sum_{k=0}^{\infty} |R_{mk}| < \infty$. Also $\sup_m E|h_x(\theta,\xi_m)|^2 < \infty$.

3. Tightness of $\{U_n\}$

In order to simplify the presentation of the chain of calculations, we present them partially in a sequence of lemmas. Among other things, we wish to show that the H_n and $\bar{h}_n$ in (2.2) can be replaced by $\bar{H}$ and h_n, resp. Apart from differences due to the greater generality of the noise here, the main differences between the treatment of (1.1) and the past work where h was assumed additive in ξ are due to the randomness of the H_n. To deal with them, we exploit the "averaging" or "smoothing" conditions (A3) and the stability condition (A4). We use K to denote a constant whose value may change from usage to usage.

Henceforth $\{\varepsilon_k\}$ denotes a sequence of positive real numbers such that $\sum_k \varepsilon_k < \infty$. Let $\{M_k\}$ be a sequence of integers tending to ∞ as $k \to \infty$, and define the measurable sets (in the sample space) A_k, B_k and C_k by (note that $j\varepsilon_k \geq t_{M_k}$ and $m(j\varepsilon_k) \geq M_k$ are equivalent statements)

$$A_k = \{ \sup_{j\varepsilon_k \geq t_{M_k}} \max_{0 \leq t \leq \varepsilon_k} \left| \sum_{i=m(j\varepsilon_k)}^{m(j\varepsilon_k+t)-1} \Delta t_i (Ah_x(\theta,\xi_i) - AH) \right| \geq \varepsilon_k^2 \},$$

$$B_k = \{ \sup_{j\varepsilon_k \geq t_{M_k}} \max_{0 \leq t \leq \varepsilon_k} \left| \sum_{i=m(j\varepsilon_k)}^{m(j\varepsilon_k+t)-1} \Delta t_i (|h_x(\theta,\xi_i)| - \tau) \right| \geq \varepsilon_k^2 \},$$

$$C_k = \sup_{j \geq M_k} \left| \int_0^1 [h_x(\theta + t(X_j - \theta), \xi_j) - h_x(\theta,\xi_j)]dt \right| \geq \varepsilon_k^2 \}.$$

Set $D_k = \bigcup_{i=k}^{\infty} (A_i \cup B_i \cup C_i)$. Choose M_k such that $P\{A_k\} + P\{B_k\} + P\{C_k\} \leq \varepsilon_k$ and $\Delta t_i \leq \varepsilon_k^2$, $i \geq M_k$. Such a choice is possible by (A3). Then $P\{D_k\} \equiv \mu_k \to 0$ as $k \to \infty$. Consequently for $\omega \notin D_k$ and $i \geq M_k$, (A3) implies that the individual terms in the sums in (A3) satisfy

$$|\Delta t_i (Ah_x(\theta,\xi_i) - AH)| \leq 4|A|\varepsilon_k^2,$$

$$|\Delta t_i (|h_x(\theta,\xi_i)| - \tau)| \leq 4\varepsilon_k^2 .$$

From the definitions of M_k and D_k we immediately get the following lemma.

Lemma 1. Under (A1) - (A3), there is a constant K such that for each k and $\omega \notin D_k$ and $j \geq M_k$,

$$\sum_{i=m(j\varepsilon_k)}^{m(j\varepsilon_k+\varepsilon_k)-1} \Delta t_i |H_i| \le K\varepsilon_k ,$$

$$\Big| \sum_{i=m(j\varepsilon_k)}^{m(j\varepsilon_k+t)-1} \Delta t_i (H_i - \overline{H}) \Big| \le K\varepsilon_k^2 , \quad t \le \varepsilon_k .$$

We now proceed to put the iteration (2.2) into a more convenient form. Define C_n^N by $C_{N+1}^N = I$ and for $n \le N$, $C_n^N = \prod_{j=n}^{N} (I+\Delta t_j H_j) \equiv (I+\Delta t_N H_N) \cdots (I+\Delta t_n H_n)$.

Lemma 2. *Assume* (A1) *to* (A3). *Then on a set whose probability is arbitrarily close to* 1

$$C_{m(t_N+s)}^{m(t_N+t+s)} \to \exp \overline{H}t \tag{3.1}$$

as $N \to \infty$, *uniformly on bounded t-intervals. Also, there is a real* K *such that for each* k *and each* $N \ge M_k$ *and* $\omega \notin D_k$ *and* $t \le \varepsilon_k$

$$C_{m(t_N+s)}^{m(t_N+t+s)} = [I + \overline{H}t + \sigma], \tag{3.2}$$

where $|\sigma| \le K\varepsilon_k^2$.

Proof. (3.1) follows directly from (3.2) and we only prove (3.2) for $t \le \varepsilon_k$ and $s = 0$. For $M \ge m$ we have

$$C_m^M = \prod_m^M (I + \Delta t_i H_i) = I + \sum_{i=m}^{M} \Delta t_i H_i + \sum_{i_2=m}^{M} \sum_{i_1>i_2}^{M} \Delta t_{i_1} \Delta t_{i_2} H_{i_1} H_{i_2} + \cdots$$

$$+ \Delta t_M \ldots \Delta t_m H_M \ldots H_m .$$

$$|C_m^M - (I+\sum_{i=m}^{M} \Delta t_i H_i)| \le \sum_{i_2=m}^{M} \sum_{i_1>i_2}^{M} \Delta t_{i_1} \Delta t_{i_2} |H_{i_1}||H_{i_2}| + \cdots + \Delta t_M \ldots \Delta t_m |H_M| \ldots |H_m| \tag{3.3}$$

$$\leq \frac{1}{2}\left(\sum_{i=m}^{M} \Delta t_i |H_i|\right)^2 + \ldots \quad .$$

Now using Lemma 1 to upper bound the right side of (3.3) and to estimate $\sum_{i=m}^{M} \Delta t_i H_i$ yields (3.2). Q.E.D.

We require one more preparatory setup. For any M, m and vector z_0 define

$$z_1 = \prod_m^M (I+\Delta t_i H_i) z_0 = C_m^M z_0,$$

where $t_{M+1}-t_m \leq \varepsilon_k$ and $m \geq M_k$. Let P denote the unique (under (A4)) symmetric positi definite matrix such that $\bar{H}'P + P\bar{H} = -I$; $x'Px$ is a Liapunov function for the differential equation $\dot{x} = \bar{H}x$, which is asymptotically stable under (A4). Define $|x|_P = (x'Px)^{1/2}$, and let u denote a <u>positive</u> constant such that $u|x|_P^2 \leq |x|^2$. By Lemma 2, if $m \geq M_k$ and $(t_{M+1}-t_m) \leq \varepsilon_k$ and $\omega \notin D_k$, we have

$$z_1 = [I + (t_{M+1}-t_m)\bar{H} + \sigma] z_0$$

where $|\sigma| \leq K\varepsilon_k^2$ and (under (A4) and using $\bar{H}'P+P\bar{H} = -I$)

$$z_1'Pz_1 = z_0'Pz_0 - (t_{M+1}-t_m)\,|z_0|^2$$

$$+ z_0'[P\sigma+\sigma'P+\sigma'P\sigma + (t_{M+1}-t_m)(\bar{H}'P\sigma+\sigma'P\bar{H}) + (t_{M+1}-t_m)^2\bar{H}'P\bar{H}]z_0$$

from which we get (for some real K)

$$\text{(3.4)} \quad |z_1|_P^2 \leq (1 - u(t_{M+1}-t_m) + K\varepsilon_k^2)|z_0|_P^2$$

$$\leq \exp[-u(t_{M+1}-t_m) + K\varepsilon_k^2]|z_0|_P^2 \; .$$

Thus $|C_m^M|_P \leq \exp[-u(t_{M+1}-t_m) + K\varepsilon_k^2]$. We are now ready for the first theorem.

<u>Theorem 1</u>. <u>Under</u> (A1) <u>to</u> (A5), $\{U_n\}$ <u>is tight on</u> R^r.

<u>Proof</u>. By iterating (2.2) we get

$$U_{N+n+1} = C_N^{N+n} U_N + \sum_{\ell=0}^{n} C_{N+\ell+1}^{N+n} A\delta\overline{W}_{n+\ell}. \tag{3.5}$$

Define

$$\overline{W}_j^m = \delta\overline{W}_j + \dots + \delta\overline{W}_m,$$

$$W_j^m = \delta W_j + \dots + \delta W_m.$$

Then a summation by parts of (3.5) yields

$$U_{N+n+1} = C_N^{N+n} U_N + C_{N+1}^{N+n} A\overline{W}_N^{N+n} - \sum_{\ell=1}^{n} C_{N+\ell+1}^{N+n} H_{N+\ell} A\overline{W}_{N+\ell}^{N+n} \Delta t_{N+\ell}. \tag{3.6}$$

The estimate (3.4) will now be used heavily. By dividing the interval $[t_N, t_{N+n+1}]$ into subintervals of length ε_k (except for the last subinterval, which is $\le \varepsilon_k$) and using (3.4), we get that there is a sequence of real numbers $\delta_k \to 0$ such that if $\omega \notin D_k$ and $N \ge M_k$, then

$$\begin{aligned} |U_{N+n+1}|_P \le\ & (1+\delta_k) \exp\left[-\frac{u}{2}(t_{N+n+1}-t_N)\right] \cdot |u_N|_P \\ & + (1+\delta_k) \exp\left[-\frac{u}{2}(t_{N+n+1}-t_N)\right] |A\overline{W}_N^{N+n}|_P \\ & + (1+\delta_k) \sum_{\ell=1}^{n} \exp\left[-\frac{u}{2}(t_{N+n+1}-t_{N+\ell})\right] \cdot \Delta t_{N+\ell} |H_{N+\ell} A\overline{W}_{N+\ell}^{N+n}|_P. \end{aligned} \tag{3.7}$$

Henceforth, <u>purely for notational convenience</u>, we suppose that the $\delta\overline{W}_i$ are scalar-valued. In general, we need only work with one component at a time anyway. Proceeding, let us next evaluate $E|W_m^M|^2$:

$$(3.8)\qquad E|W_m^M|^2 = E\sum_{i,j=m}^{M}\sqrt{\Delta t_i}\sqrt{\Delta t_j}\,h_i h_j \le 2\sum_{i=m}^{M}\sqrt{\Delta t_i}\sum_{j\ge i}^{M}\sqrt{\Delta t_j}\,|Eh_i h_j|$$

$$\le 2\sum_{i=m}^{M}\sqrt{\Delta t_i}\sum_{j\ge i}\sqrt{\Delta t_j}\,|R_{ij}| \le 2K\sum_{i=m}^{M}\Delta t_i = 2K(t_{M+1}-t_m),$$

where the last inequality follows by the first half of (A5). With perhaps a different K, the same inequality holds for $E|\overline{W}_m^M|^2$. By this estimate and the second half of (A5), there is a constant K_k such that for $N \ge M_k$

$$(3.9)\qquad E|H_{N+\ell}A\overline{W}_{N+\ell}^{N+n}|_P\, I_{\{\omega\notin D_k\}} \le K_k(t_{N+n+1}-t_{N+\ell})^{1/2}.$$

Inequality (3.7) holds with probability $1 - P\{D_k\} = \rho_k \to 1$. Let us modify the $\{U_i, H_i, i\ge M_k\}$ on D_k in a way such that (3.7) holds for all n and (3.9) holds without the indicator function and where K_k does not depend on k. Let $\{U_i^k, H_i^k\}$ denote the altered sequence. Then (3.7) and (3.9) together imply that $\sup_{i\ge M_k} E|U_i^k|^2 < \infty$. Thus the sequence $\{U_i, i<M_k;\ U_i^k, i\ge M_k\}$ is tight on R^r. Since k is arbitrary and $\rho_k \to 1$ as $k\to\infty$, this implies that the original $\{U_i\}$ sequence is tight. Q.E.D.

4. Weak Convergence of $\{U^N(\cdot)\}$ and the Rate of Convergence

In this section, we show that $\{U^N(\cdot)\}$ converges weakly in $D^r[0,\infty)$ to the stationary solution to the Gauss-Markov diffusion

$$(4.1)\qquad dU = \overline{H}U dt + AR^{1/2}dB,$$

where $B(\cdot)$ is a standard Wiener process and $R^{1/2}$ is a square root of the matrix R in (A6) below. In particular, this implies that $(X_n-\theta)(n+1)^{\alpha/2}$ converges in distribution to a normal random variable with mean 0 and covariance

$$\int_0^\infty (\exp \bar{H}t)ARA'(\exp \bar{H}'t)dt.$$

We will require the following additional assumptions.

(A6) $\{h_j\}$ is a stationary sequence, and $E|h_j|^6 < \infty$. Define $R(i) = Eh_j h'_{j+1}$. Then $R \equiv \sum_{-\infty}^{\infty} R(i)$ is bounded by (A8).

Let $\mathscr{B}_j = \mathscr{B}(h_\ell, \ell \le j)$ and let E_j denote the expectation conditional on $\mathscr{B}_j$.

(A7) Define $\rho_1(i)$ by

$$\rho_1(i) = \sup_{j,\ell \ge 0} E^{1/2}|E_j h_{j+i} h'_{j+i+\ell} - R(\ell)|^2.$$

Then $\sum_i \rho_1^{1/2}(i) < \infty$.

The $\sup_j$ above and $\sup_k$ below are redundant if we assume that the $\{h_j\}$ process started at $j = -\infty$, and choose the sample space appropriately.

(A8) Define $\rho_2(i)$ by $\rho_2(i) = \sup_k E^{1/2}|E_k h_{k+i}|^2$. Then $\sum_i \rho_2^{1/2}(i) < \infty$.

We now give some examples of (A7) and (A8). First suppose that $\{h_j\}$ is a stationary and bounded ϕ-mixing process in the sense of [7, p. 166], with of course $Eh_j \equiv 0$. Let K denote an arbitrary constant. By [8, Lemma 1], $|E_j h_{j+k}| \le K\phi_k$ and $|E_j h_{j+k} h'_{j+k+\ell} - R(\ell)| \le K\phi_k$. Thus $\rho_1(i) \le K\phi_i$, $\rho_2(i) \le K\phi_i$. If $\sum_\ell \phi_\ell^{1/2} < \infty$, then (A7) and (A8) hold. However, if the h_j are bounded and ϕ-mixing, then a slightly different proof of Theorem 2 can be given, requiring only $\sum_\ell \phi_\ell^{1/2} < \infty$.

An example of (A6) to (A8). Let Q denote a matrix whose eigenvalues are inside the unit circle, let $\{\psi_n\}$ denote a sequence of independent and identically distributed Gaussian random variables and define ξ_n, $\infty > n > -\infty$, by $\xi_{n+1} = Q\xi_n + \psi_n$. Then $\{\xi_n\}$ is a stationary sequence. Let $Eh(\theta,\xi_j) \equiv Eh_j = 0$ and suppose that $\bar{h}(\cdot) = h(\theta,\cdot)$ satisfies a uniform Lipschitz condition, with constant K_1. Let $\mathcal{G}_j$ measure ψ_i, $i \leq j$.

Let us evaluate $E|E_k\bar{h}(\xi_{k+i})|$. Let $\{\tilde{\psi}_i\}$ denote a sequence with the same distribution as $\{\psi_i\}$, but independent of it. We have

$$\xi_{k+i} = Q^i\xi_k + \sum_{\ell=0}^{i-1} Q^\ell \psi_{k+i-\ell-1}$$

which has the same distribution as

$$\sum_{\ell=0}^{\infty} Q^\ell\tilde{\psi}_\ell - \sum_{\ell=i}^{\infty} Q^\ell\tilde{\psi}_\ell + Q^i\xi_k$$

Using the fact that the first term above has the same distribution as ξ_m has for any m, together with the Lipschitz condition, yields

$$|E[\bar{h}(\text{first term} - \sum_{\ell=1}^{\infty} Q^\ell\tilde{\psi}_\ell + Q^i\xi_k) - E\bar{h}(\text{first term})|\xi_k]| \leq K_1E|\sum_{\ell=1}^{\infty} Q^\ell\tilde{\psi}_\ell| + K_1|Q^i\xi_k|.$$

from which (A8) follows. A similar (and omitted) calculation yields (A7).

<u>Theorem 2</u>. <u>Under</u> (A1) - (A8), $\{U^N(\cdot)\}$ <u>converges weakly to the stationary solution to</u> (4.1).

<u>Part 1</u>. Define the "approximation to a Wiener process" $W^N(\cdot)$ by

$$W^N(t) = W_N^{m(t_N+t)-1} = \sum_{i=N}^{m(t_N+t)-1} \sqrt{\Delta t_i}\, h_i,$$

with a similar definition for $\overline{W}^N(\cdot)$ (but using $\delta\overline{W}_i$ in lieu of δW_i). We will show that $\{W^N(\cdot)\}$ is tight in $D^r[0,\infty)$ and converges to a Wiener process with covariance matrix Rt. It easily follows from this that the same result must hold for $\{\overline{W}^N(\cdot)\}$, since $(n+2/n+1)^{\alpha/2} = 1 + O(\frac{1}{n})$ implies that $\{|W^N(\cdot)-\overline{W}^N(\cdot)|\}$ tends weakly to the zero process.

First we prove <u>tightness</u> of $\{W^N(\cdot)\}$. For notational convenience only, we assume that the h_j are scalar-valued in this part of the proof. Otherwise, we would work with one component at a time anyway, so there is no loss of generality.

Let $\ell \geq k \geq j \geq i$. We have

$$(4.2)\quad |Eh_ih_jh_kh_\ell| \leq |Eh_ih_jh_kh_\ell - Eh_ih_jEh_kh_\ell| + |Eh_ih_j|\ |Eh_kh_\ell| .$$

The first term on the right satisfies (use (A7))

$$|Eh_ih_j(E_jh_kh_\ell - Eh_kh_\ell)| \leq E^{1/2}|h_ih_j|^2E^{1/2}|E_jh_kh_\ell - Eh_kh_\ell|^2 \leq K\rho_1(k-j).$$

By (A8), the first term on the right of (4.2) is bounded above by

$$|Eh_ih_jh_kE_kh_\ell| + |Eh_ih_jEh_kE_kh_\ell| \leq E^{1/2}|h_ih_jh_k|^2E^{1/2}|E_kh_\ell|^2$$

$$+ |Eh_ih_j|E^{1/2}h_k^2E^{1/2}|E_kh_\ell|^2$$

$$\leq K\rho_2(\ell-k).$$

Thus

$$(4.3)\quad |Eh_ih_jh_kh_\ell| \leq K\rho_1^{1/2}(k-j)\rho_2^{1/2}(\ell-k) + |R(j-i)|\ |R(\ell-k)|.$$

Using these bounds, we get

$$E|W^N(t+s)-W^N(t)|^4 = E\left|\sum_{i=m(t_N+t)}^{m(t_N+t+s)-1} \sqrt{\Delta t_i}\ h_i\right|^4$$

$$\leq K \sum_{i\leq j\leq k\leq \ell} (\Delta t_i\Delta t_j\Delta t_k\Delta t_\ell)^{1/2}|Eh_ih_jh_kh_\ell|$$

(summation between $m(t_N+t)$ and $m(t_N+t+s)-1$; at each use of K it may have a different value)

$$\leq K \sum_{i\leq j\leq k\leq \ell} (\Delta t_i \Delta t_j \Delta t_k \Delta t_\ell)^{1/2}[\rho_1^{1/2}(k-j)\rho_2^{1/2}(\ell-k) + |R(j-i)|\cdot|R(\ell-k)|],$$

(sum over ℓ and use $\Delta t_k \geq \Delta t_\ell$)

$$\leq K \sum_{i\leq j\leq k} (\Delta t_i \Delta t_j)^{1/2} \Delta t_k [\rho_1^{1/2}(k-j) + |R(j-i)|]$$

(sum over j and use $\Delta t_i \geq \Delta t_j$)

$$\leq K \sum_{i\leq k} \Delta t_i \Delta t_k \leq Ks^2 \tag{4.3}$$

where the last inequality holds if t_N+t+s and t_N+t take values in the set $\{t_i\}$.

If (4.3) holds for all t, s, N, then [7], Theorems 15.5 and 12.3 imply that $\{W^N(\cdot)\}$ is tight in $D^r[0,\infty)$ and that all processes which are weak limits have continuous paths w.p. 1. But, since $\Delta t_n \to 0$ and the paths are piecewise constant, it is enough that (4.3) hold for t_N+t+s and t_N+t in the $\{t_i\}$ set. Thus $\{W^N(\cdot)\}$ is tight and all limit processes have continuous paths w.p. 1.

Part 2. Now, the h_i are treated as vectors rather than scalars. Let N index a weakly convergent subsequence of $\{W^N(\cdot)\}$ and denote the (continuous w.p. 1) weak limit by $W(\cdot)$. Note that (4.3) implies that $\{|W^N(\cdot)|^2\}$ is uniformly integrable. Let $s_i \leq t \leq t+s$ and q be arbitrary. Let $g(\cdot)$ denote a bounded continuous function of $W^N(s_i)$, $i \leq q$, and let E_t^N denote expectation conditioned on $\{h_j, j\leq m(t_N+t)-1\}$. Then

$$Eg(W^N(s_i), i\leq q)[W^N(t+s)-W^N(t)]$$

$$= Eg(W^N(s_i), i\leq q)E_t^N \sum_{i=m(t_N+t)}^{i=m(t_N+t+s)-1} \sqrt{\Delta t_i}\, h_i$$

goes to zero as $N \to \infty$ by (A8). This together with the uniform integrability and weak convergence imply that $Eg(W(s_i), i\leq q)[W(t+s)-W(t)] = 0$ for all q, bounded continuous g and $\{s_i\} \leq t \leq t+s$. Thus $W(\cdot)$ is a continuous martingale. To compute its quadratic variation, repeat the above argument with $[W^N(t+s)-W^N(t)]$ $[W^N(t+s)-W^N(t)]'$ replacing $[W^N(t+s)-W^N(t)]$. Using (A6), the weak convergence and uniform integrability yields

$$Eg(W^N(s_i),\ i\leq q)[W^N(t+s)-W^N(t)][W^N(t+s)-W^N(t)]' \to Eg(W(s_i),\ i\leq q)\ Rs.$$

Then the arbitrariness of g and $s_i \leq t \leq t+s$ yield that the quadratic variation (at s) is Rs. Thus $W(\cdot)$ is a Wiener process with covariance Rs, as asserted. This result does not depend on the chosen convergent subsequence.

Part 3. Define the function $C^n(t,t+s) = C_{m(t_N+t)}^{m(t_N+t+s)-1}$. Define a function $H^N(\cdot)$ with values $H_t^N = H_{N+n}$ in $[t_{N+n}-t_N, t_{n+N+1}-t_N)$.

Then for $t \in \{t_{N+i}-t_N,\ i\geq 0\}$, and modulo a factor for each term which goes to zero uniformly in t w.p. 1 as $N \to \infty$, the sum (3.6) can be written in the integral form (since the integrand is constant over Δt_i intervals)

$$U^N(t) = C^N(0,t)U^N(0) + C^N(0,t)A\overline{W}^N(t) - \int_0^t C^N(s,t)H_s^N A[\overline{W}^N(t)-\overline{W}^N(s)]ds. \tag{4.4}$$

for $t > 0$, between the $\{t_i\}$, the integral in (4.4) is just a linear interpolation instead of a piecewise constant interpolation of the sum in (3.6), and we may work with it instead. Define $H^N(\cdot)$ by $N^N(t) = \sum_{i=m(t_N)}^{m(t_N+t)-1} H_i\Delta t_i$. By (A3), $\{H^N(\cdot)\}$ is tight in $D^r[0,\infty)$ and all limits are the constant process with value $\overline{H}t$ at t. Note that $\{C^N(0,t)\}$ is tight on $D^q[0,\infty)$ for an appropriate integer q, since it converges to $\exp \overline{H}t$ uniformly on bounded intervals w.p. 1.

We now have essentially all the limits that are required. If H_s^N converged to the constant $\overline{H}$ w.p. 1 as $N \to \infty$, then the weak convergence of $\overline{W}^N(\cdot)$ and convergence of $C^N(s,t)$ would imply that (4.4) holds with all functions replaced by their limits (and a weakly convergent subsequence of $\{U^N(0)\}$ taken). Since H_s^N does not usually converge in the above sense, a slightly indirect method must be used to allow us to make the replacements suggested above. It is convenient to have all the random functions defined on the same space and to work with w.p. 1 rather than with weak convergence. To do this we apply the imbedding technique of Skorokhod [9], Theorem 3.1.1. The family $\{U^N(0), H^N(\cdot), \overline{W}^N(\cdot), C^N(0,t)\} \equiv \{\Phi^N(\cdot)\}$ is tight in the appropriate space $R^r \times D^{2r+q}[0,\infty) \equiv \mathcal{D}$ and all limit functions are continuous w.p. 1. Extract a convergent subsequence, index it by N, and denote the limit by $(U(0), \overline{H}(\cdot), W(\cdot), C(0,\cdot) \equiv \Phi(\cdot)$. By the Skorokhod imbedding method [9], Theorem 3.1.1, there exists a probability space $(\tilde{\Omega}, \tilde{P}, \tilde{B})$ with random processes $\{\tilde{U}^N(0), \tilde{H}^N(\cdot), \tilde{W}^N(\cdot), \tilde{C}^N(0,\cdot)\} \equiv \{\tilde{\Phi}^N(\cdot)\}$ and $(\tilde{U}(0), \tilde{H}(\cdot), \tilde{W}(\cdot), \tilde{C}(0,\cdot)) \equiv \tilde{\Phi}(\cdot)$ defined on it, where $\tilde{\Phi}^N(\cdot)$ (resp., $\tilde{\Phi}(\cdot)$) has the same distribution as $\Phi^N(\cdot)$ (resp., $\Phi(\cdot)$), all the processes in $\tilde{\Phi}(\cdot)$ have continuous paths and $\tilde{\Phi}^N(\cdot) \to \tilde{\Phi}(\cdot)$ w.p. 1 in the topology of $\mathcal{D}$. Since the limit processes are continuous, this means uniform convergence on bounded intervals. From $\tilde{H}^N(\cdot)$, we can recover the random variables $\tilde{H}_{N+i}$, $i \geq 0$, from which it was constructed, since $\tilde{H}^N(\cdot)$ is also piecewise constant w.p. 1. Also $\{\tilde{H}_{N+i}, i\geq 0\}$ has the same distribution as has $\{H_{N+i}, i\geq 0\}$.

We work with the imbedded processes, but drop the tilde affix. Now, return to (4.4) and, via the imbedding, suppose that all weak convergences are w.p. 1 in the above-cited topology. The first two terms of (4.4) converge to $(\exp \overline{H}t)\, U(0)$ and $(\exp \overline{H}t)\, W(t)$, resp. Note that $C^N(s,t) = C^N(0,t)[C^N(0,s)]^{-1}$ also converges w.p. 1 uniformly on bounded sets to $\exp \overline{H}(t-s)$. We next write the integral in (4.4) in a more convenient way.

Let $\Delta > 0$, and let $M = \max\{i: i\Delta \leq t\}$. We have

$$\sum_{i=0}^{M-1} \Big| \int_{i\Delta}^{i\Delta+\Delta} \{C^N(s,t)H_s^N A[W^N(t)-W^N(s)] - C(i\Delta,t)H_s^N A[W(t)-W(i\Delta)]\}ds$$

$$+ \int_{M\Delta}^{t} C^N(s,t)H_s^N A[W^N(t)-W^N(s)] - C(M\Delta,t)H_s^N A[W(t)-W(M\Delta)]\}ds$$

$$\leq \sum_{i=0}^{M-1} \sup_{i\Delta \leq s \leq i\Delta+\Delta} [|C^N(s,t)-C(i\Delta,t)| + |W^N(s)-W(i\Delta)| + |W(t)-W^N(t)|]$$

$$\cdot [|W^N(t)-W^N(s)| + |C(s,t)|] \int_{i\Delta}^{i\Delta+\Delta} |H_s^N| ds |A| \text{ plus a similar expression for the end term.}$$

By the w.p. 1 uniform convergences (on bounded intervals) and continuity of the limit functions and the estimate (A3b), the limit of the above expression goes to zero uniformly on bounded t sets, w.p. 1, as $N \to \infty$ and then $\Delta \to 0$.

Thus, we need only examine the limits of

$$(4.5) \qquad \sum_{i=0}^{M-1} \int_{i\Delta}^{i\Delta+\Delta} C(i\Delta,t)H_s^N A[W(t)-W(i\Delta)]ds + \int_{M\Delta}^{t} C(M\Delta,t)H_s^N A[W(t)-W(M\Delta)]ds.$$

But, by (A3a), (4.5) converges to the same expression with $\overline{H}$ replacing H_s^N, uniformly on bounded intervals, w.p. 1 as $N \to \infty$. By the above calculations we can write the limit of the third term in (4.4) as

$$(4.6) \qquad -\int_0^t C(s,t)\overline{H}A[W(t)-W(s)]ds$$

for the imbedded, hence the original processes. Thus $U^N(t)$ (the imbedded process) converges to

$$(4.7) \qquad U(t) \equiv C(0,t)U(0) + C(0,t)A\,W(t) + (4.6)$$

uniformly on finite intervals, w.p. 1. Consequently the original $U^N(\cdot)$ converges weakly to the process (4.7). But (4.7) is the unique solution to (4.1) with initial condition U(0). The form is independent of the selected convergent subsequences. Also, via an integration by parts,

$$U(t) = C(0,t)\,U(0) + \int_0^t C(s,t)\,A\,dW_s. \tag{4.8}$$

We need only show that U(0) is the "stationary" initial condition. This can be easily shown in the following manner. The set of all possible U(0) is tight because $\{U_n\}$ is. Also the weak limits of $\{U^N(\cdot)\}$ are also weak limits of the restrictions to $T,\infty)$ of the weak limits of (the functions are left-shifted by T) $\{U^{m(t_N-T)}(\cdot)\}$ on $D^r[0,\infty)$, since $U^{m(t_N-T)}(T) = U_N$. But the latter limits are of the form (4.8) also. The restriction to $[T,\infty)$ involves simply replacing t by T+t in (4.8). From this, the tightness of possible U(0), the arbitrariness of T and the fact that $C(0,t) = \exp \bar{H}t \to 0$ as $t \to \infty$, we get that U(0) must be the "stationary" initial condition. Q.E.D.

References

1. J. Sacks, "Asymptotic distribution of stochastic approximation procedures", Ann. Math. Statist. 29 (1958), pp. 273-405.

2. V. Fabian, "On asymptotic normality in stochastic approximation", Ann. Math. Statist. 39 (1968), pp. 1327-1332.

3. H.J. Kushner, "Rates of convergence for sequential Monte-Carlo optimization methods", SIAM J. on Control and Optimiz. 16 (1978), pp. 150-168.

4. H.J. Kushner and D.S. Clark, Stochastic approximation methods for constrained and unconstrained systems, Applied Math. Sci. Series no. 26 (1978), Springer, Berlin.

5. L. Ljung, "Analysis of recursive stochastic algorithms", IEEE Trans. on Automatic Control AC-22 (1977), pp. 551-575.

6. L. Ljung, "On positive real transfer functions and the convergence of some recursive schemes", IEEE Trans. on Automatic Control AC-22 (1977), pp. 539-550.

7. P. Billingsley, Convergence of probability measures, Wiley (1968) New York.

8. G.C. Papanicolaou and W. Kohler, "Asymptotic theory of mixing stochastic processes", Comm. Pure and Appl. Math. 27 (1974), pp. 641-668.

9. A.V. Skorokhod, Limit theorems for stochastic processes, Theory of probability and its applications, 1 (1956), pp. 262-290.